## Raport z ćwiczenia ML[[1]](#footnote-1)

|  |
| --- |
| Data: 18.04 |
| Imię i nazwisko: Maciej Adamus |

Sprawozdanie z ćwiczeń laboratoryjnych powinno składać się z TRZECH części (chyba instrukcja do ćwiczenia określa to inaczej).

* REZULTATY

*Zanotuj określone w treści ćwiczenia parametry algorytmów, otrzymane rezultaty, itp.   
Opc. zamieść listę dodatkowych plików dołączonych do sprawozdania (dodatkowe pliki to np. fragmenty kodu, pliki danych otrzymane w trakcie ćwiczenia, itp.)*

* ANALIZA i WNIOSKI  
  *Zamieść, określone w treści ćwiczenia, analizę otrzymanych rezultatów (np. statystyczne opracowanie wyników) oraz wnioski. Maksymalnie 1 strona.*
* ODPOWIEDZI NA PYTANIA  
  *Zamieść, określone w treści ćwiczenia, odpowiedzi na pytania. Maksymalnie 1 strona.*

Spis treści

[***Raport z ćwiczenia ML*** 1](#_Toc164938869)

[Rezultaty 2](#_Toc164938870)

[Analiza i wnioski 5](#_Toc164938871)

[Odpowiedzi na pytania 7](#_Toc164938872)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rezultaty  1. Liczebności klas   Liczebność danych z każdej klasy jest jednakowa i wynosi 12. Równowaga w liczebności klas jest ważna, ponieważ wtedy klasyfikator może w miarę równomiernie nauczyć się cech dotyczących każdej z klas.   1. Podział na zbiór uczący i testowy   Liczebności poszczególnych klas po podziale na zbiory uczący i testowy są równe dla każdej klasy i wynoszą 8 / 4.  Liczebności poszczególnych klas po podziale na zbiory uczący i testowy przy użyciu **cvpartition** prezentują się następująco:   |  |  |  | | --- | --- | --- | | Samogłoska | Ilość danych uczących | Ilość danych testowych | | A | 9 | 3 | | E | 8 | 4 | | I | 9 | 3 | | O | 8 | 4 | | U | 8 | 4 | | Y | 9 | 3 |  1. Dwa najlepsze klasyfikatory 2. Weighted KNN: accuracy 72.9%   Obraz zawierający zrzut ekranu, kwadrat, Prostokąt, diagram  Opis wygenerowany automatycznie   1. Fine KNN: accuracy 64.6%      1. PCA on Fine KNN: accuracy 77.1%      1. Wyniki najlepszego klasyfikatora na danych testowych   Accuracy: 79.17%  Macierz pomyłek:  4 0 0 0 0 0  0 4 0 0 0 0  0 0 3 0 1 0  0 0 0 2 2 0  0 0 0 0 3 1  0 1 0 0 0 3 |

|  |
| --- |
| Analiza i wnioski  1. Czy liczba wczytanych elementów klas jest podobna do siebie? Wyjaśnij dlaczego to ma znaczenie w kontekście uczenia maszynowego.   - Liczebność danych z każdej klasy jest jednakowa i wynosi 12. Równowaga w liczebności klas jest ważna, ponieważ wtedy klasyfikator może w miarę równomiernie nauczyć się cech dotyczących każdej z klas.   1. Czy w przypadku prostej walidacji za każdym razem liczebności grup są takie same? Jeśli nie to wyjaśnij dlaczego.   - Gdy ręcznie dzielimy zbiory to liczebności są takie same. Kiedy używamy cvpartition to rozmiary zbiorów delikatnie się różnią.   1. Wyjaśnij na czym polegają: prosta walidacja (holdout validation) oraz walidacja krzyżowa (k-fold cross-validation) i jaki jest jej cel. Jakie są inne sposoby podziału zbioru danych na zbiór uczący i testowy? Jaką walidację powinno się stosować w przypadku gdy zbiór danych jest niewielki?   - Prosta walidacja polega na podziale zioru na podzbiór uczący i testowy, trenowanie modelu na danych uczących a następnie zweryfikowanie go na danych których nigdy nie widział.  K-krotna walidacja krzyżowa polega na podziale zbioru na k części a następnie wytrenowanie go k razy, za każdym razem korzystając z innego zbioru jako danych testowych. Celem tego jest dokładniejsza ocena skuteczności modelu.  Inne sposoby podziału zbioru danych na zbiór uczący i testowy to np. leave-one-out cross-validation (LOOCV).  W przypadku niewielkich zbiorów danych zalecane jest stosowanie walidacji krzyżowej.   1. Na podstawie elementów w macierzy pomyłek zanotuj które samogłoski są ze sobą najbardziej mylone.   - Najczęściej mylone są pary a-e oraz o-u.   1. Wyjaśnij na czym polega specyficzność oraz precyzja klasyfikatora oraz pojęcia „True Positive Rates”, „False Negative Rates”.   Specyficzność - zdolność do poprawnego identyfikowania negatywnych przypadków spośród wszystkich negatywnych przypadków w zbiorze danych.  Precyzja - miara skuteczności w identyfikowaniu pozytywnych przypadków spośród wszystkich przypadków, które klasyfikator zidentyfikował jako pozytywne.  True Positive Rate - stosunek poprawnie zidentyfikowanych pozytywnych przypadków do wszystkich rzeczywiście pozytywnych przypadków w zbiorze danych.  False Negative Rate - odsetek przypadków, które zostały błędnie sklasyfikowane jako negatywne, podczas gdy są one w rzeczywistości pozytywne.   1. Jaka jest rola PCA (ang. Principal Component Analysis) w klasyfikacji?   - Głównym celem jest redukcja wymiarowości poprzez transformację przestrzeni, zachowując przy tym jednocześnie jak najwięcej informacji.   1. Dlaczego stosujemy 3 zbiory: uczący, walidacyjny oraz testowy?   - Zbiór uczący służy do trenowania modelu, zbiór walidacyjny jest używany do oceny wydajności modelu w trakcie procesu trenowania, zbiór testowy do oceny modelu   1. Wyjaśnij dlaczego rezultaty (validationAccuracy) zwracane przez wygenerowaną funkcję różnią się od rezultatów dla całego zbioru uczącego (accuracyTrain)?   - Ta róznica wynika z tego, że accuracyTrain mierzy wydajność modelu na danych treningowych, a validationAccuracy ocenia jego zdolność do generalizacji na nowe dane poprzez ocenę na zbiorze walidacyjnym, który nie był używany w procesie uczenia. |

|  |
| --- |
| Odpowiedzi na pytania  1. Wymień i krótko uzasadnij 3 potencjalne praktyczne zastosowania algorytmów przedstawionych w ćwiczeniu.    1. Klasyfikacja spamu w skrzynce pocztowekj – automatyczna ocena maili pod kątem spamu    2. Rozpoznawanie tekstu – automatyczne wykrywanie tekstu na obrazach    3. System rekomendacji w sklepach internetowych – na podstawie zachowań klientów mogą polecać podobne przedmioty 2. Wymień i krótko uzasadnij 2 elementy z ćwiczenia, które zwróciły Twoją uwagę lub są Twoim zdaniem istotne do zrozumienia działania opisanych algorytmów.    1. Macierz konfuzji pozwala dobrze zrozumieć jak model dopasował się do danych i znaleźć ewentualne pole do poprawy    2. Walidacja krzyżowa to dobry sposób do weryfikacji modelu 3. Zadaj 1 pytanie związane z ww przykładami - co jest dla Ciebie niezrozumiałe lub niejasne.   - dlaczego cvpartition nie zawsze dzieli zbiory w równych proporcjach |

1. *Raport z ćwiczenia należy dostarczyć poprzez system UPEL, w formacie PDF.* [↑](#footnote-ref-1)